The Behavior Analyst 1999, 22, 109–116 No. 2 (Fall)

Statistical Inference in Behavior Analysis: Experimental Control is Better

Michael Perone
West Virginia University

Statistical inference promises automatic, objective, reliable assessments of data, independent of the skills or biases of the investigator, whereas the single-subject methods favored by behavior analysts often are said to rely too much on the investigator's subjective impressions, particularly in the visual analysis of data. In fact, conventional statistical methods are difficult to apply correctly, even by experts, and the underlying logic of null-hypothesis testing has drawn criticism since its inception. By comparison, single-subject methods foster direct, continuous interaction between investigator and subject and development of strong forms of experimental control that obviate the need for statistical inference. Treatment effects are demonstrated in experimental designs that incorporate replication within and between subjects, and the visual analysis of data is adequate when integrated into such designs. Thus, single-subject methods are ideal for shaping—and maintaining—the kind of experimental practices that will ensure the continued success of behavior analysis.

Science is a social enterprise, and the standards of scientific evidence are established by consensus. From this perspective, the objective of research design and data analysis is straightforward: to convince an audience of skeptical colleagues that a particular interpretation or inference is justified. The rules of statistical inference, set forth in classic texts and promulgated in mandatory graduate courses, provide an agreed-upon solution. By following these rules—and rejecting the null hypothesis with a p value of less than .05—investigators assure their peers, and themselves, of the significance of their findings. Statistics guide investigators to inferences about their data that can be expressed in objective, quantitative terms. Indeed, the inferences seem to arise automatically from the application of the statistical formula, as implied by the term most commonly used to describe the process: *statistical* inference. When scientific inferences are produced by a formula, the *investigator* is relieved of a burdensome responsibility, and science itself is protected from the frailties of human judgment, which is error-prone and subject to an assortment of troubling biases.

Requests for reprints should be sent to Michael Perone at the Department of Psychology, West Virginia University, P.O. Box 6040, Morgantown, West Virginia 26506-6040 (E-mail: mperone@wvu.edu).

Limitations of Statistical Inference

Or so it seems. Unfortunately, statistics offer the investigator no panacea, and no self-respecting statistician would claim otherwise. Despite the central role played by null-hypothesis statistical tests throughout the biological, behavioral, and social sciences, fundamental problems have been recognized for some time (e.g., Bakan, 1966; Lykken, 1968; Meehl, 1967). By 1970, the criticisms of statistical inference had drawn enough attention from psychologists to warrant a book provocatively entitled *The Significance Test Controversy* (Morrison & Henkel, 1970). But actual use of statistical analysis has not changed much since then; null-hypothesis testing is as robust as ever. Cautions may be decreed by textbook authors and professors in statistics courses, but when students and investigators are confronted with real research problems, they are beguiled by the reassuring directness of statistical procedures, which offer simple rules for answering a host of practical questions ("How many subjects per cell?"). In return, textbooks and professors seem more than willing to

offer simple recipes for cooking up the answers ("Run Cohen's power analysis program and see what it says"). Browse through an assortment of statistics texts, and you will find many with handy tables and flow-charts to guide the reader to just the right test for the data at hand, putting the task of analyzing the results of an experiment on the same level as looking up a telephone number. The appearance of "point-and-click" software for statistical analysis has made matters worse. When asked by a puzzled associate editor to explain an unusual statistic in a manuscript submitted for publication, more than one author has responded by providing the name and version number of the software package.

The bottom line is this: Too much research design and data analysis is performed without thinking. Thompson (1998) voiced his objection this way:

> Most researchers mindlessly test [the null hypothesis] because most statistical packages only test such hypotheses. This . . . does not require researchers to thoughtfully extrapolate expected results from the previous literature or from theory. Instead, science becomes an automated, blind search for mindless tabular asterisks using thoughtless hypotheses. (p. 799)

Although statistical analysis has its defenders (e.g., Dixon, 1998; Hagen, 1997, 1998; Wilcox, 1998), the criticisms of years past continue to cause trouble, and debate about statistical strengths and weaknesses is being repeated and expanded by a new generation of psychologists and statisticians (e.g., Cohen, 1994; McGrath, 1998; Tryon, 1998). The greatest strength of statistical inference—the automatic, objective, reliable assessment of data, all independent of the skills or biases of the investigator—is a mirage. Research summarized by Tryon indicates that statistical tests are routinely misinterpreted by investigators publishing in our best journals, and even by statisticians themselves. "How much more susceptible to misinterpretation," he asks, "are the vast majority of other less well quantitatively trained psychologists?" (p. 796).

On the Search for Methodological Imperatives

Eventually some investigators discover that there is no good cookbook for delicious servings of research design and data analysis. But too many still see design and analysis as obstacles to good research rather than an integral part of it. If they have grants, they hire statistical consultants. If they are graduate students, they make sure a statistics professor is a member of their dissertation committee. The prevailing attitude is that the framing of research questions can proceed apart from the methods employed to answer them.

The attraction to formulas and rules is not confined to investigators who favor group-statistical approaches. Professors who teach courses in single-subject research design are confronted by students of behavior analysis seeking, for example, rules about the criteria used to decide that behavior has reached a steady state. Over the years many students have reported that they adopted the criterion recommended by Sidman (1960) in his classic *Tactics of Scientific Research.* But Sidman never offered such a recommendation. In answer to the question "How does one select a steady-state criterion?" he explained, "There is . . . no rule to follow, for the criterion will depend upon the phenomenon being investigated and upon the level of experimental control that can be maintained" (p. 258). On what basis, then, is one to decide? Sidman pointed to the investigator's "accumulated experience and good experimental judgment" developed in the course of "designing and carrying out steady-state experiments" (p. 261).

We are left with a dilemma: Group-statistical methods incorporate tidy sets of rules, but the rules lead to less than satisfactory results, even in the hands of veterans. Single-subject methods

seem to offer no rules at all. What guidance, then, is to be offered the student embarking on a research career?

The answer may be found in three critical notions in the passages quoted from Sidman's (1960) book: experience, experimental control, and judgment. These are recurring themes in Sidman's treatment of research tactics and, indeed, throughout the historical development of behavior analysis. Understanding the role they can play in scientific research is the key to appreciating why statistical inference has not been and need not become a major factor in the experimental analysis of behavior.

Experience

Behavior analysts' interest in single-subject as opposed to group-statistical research may be regarded as the result of an inductive process arising from intense interactions with data and various practical considerations, rather than deductions from a well-developed philosophy of science. The sophisticated philosophical justification for single-subject research came later.

Behavior-analytic methods, of course, derive from the work of Skinner, whose graduate training antedated the widespread adoption of the group-statistical approach made possible by Fisher (1925). Skinner's early research involved single-subject designs; most of the experiments reported in his seminal work, *The Behavior of Organisms* (Skinner, 1938), used only 4 rats. But as large-group methods gained favor within psychology in the late 1930s, Skinner, then an assistant professor at the University of Minnesota, gave them a try. He and Heron built a set of 24 operant chambers and cumulative recorders, interconnected so that the recorders displayed mean performances for the entire group of 24 rats, as well as subgroups of 12 and 6. Skinner said that he and Heron "thus provided for the design of experiments according to the principles of R. A. Fisher, which then were coming into vogue" (Skinner, 1956/1972, p. 113). Skinner was enthusiastic about the approach; he reported that "the possibility of using large groups of animals greatly improves upon (our) method . . . since tests of significance are provided for and properties of behavior not apparent in single cases may be more easily detected" (Skinner, 1956/1972, p. 113). But Skinner's enthusiasm soon faded:

> In actual practice that is not what happened. . . . You cannot easily make a change in the conditions of an experiment when twenty-four apparatuses have to be altered. Any gain in rigor is more than matched by a loss in flexibility. We were forced to confine ourselves to processes which could be studied with the baselines already developed in earlier work. We could not move on to the discovery of other processes or even to a more refined analysis of those we were working with. No matter how significant might be the relations we actually demonstrated, our statistical Leviathan had swum aground. (Skinner, 1956/1972, pp. 113–114)

Skinner, the consummate tinkerer, was quite willing to scout about for new ways to conduct experiments. He rejected group-statistical methods not because they collided with his radical behaviorist epistemology, but rather because his experience revealed that they insulated the investigator from the behavior of the subject. The ongoing interaction between experimenter and data that had characterized his earlier work—and led to his innovations in apparatus, measurement, and theory—could not be sustained in group-statistical research. Skinner returned to the experimental analysis of individual behavior, and directed his energies to developing stronger methods of experimental control that would obviate the need for statistical inference.

Experimental Control

The tension between group-statistical and single-subject methods is created by the relative roles played by experimental control in the two approaches. For Skinner and other advocates of single-subject research, group-statistical methods are ill suited to the development of strong forms of experimental control over behavior, in

part because the group methods are unwieldy and in part because the nature of statistical analysis reduces the investigator's motivation to establish such control. The sensitivity of a statistical test is a direct function of the number of subjects, and weak control can be tolerated if the number is large enough. Averaging data across many subjects can hide a multitude of sins: The experimental treatment may fail to affect the behavior of some subjects, and may even lead to contrary effects in others. As a consequence, statistically significant results based on large sample sizes are not persuasive. Given a sufficiently large sample, statistical significance is assured. Meehl (1967) pointed out that the only question is whether the direction of the statistical difference will support the investigator's hypothesis. Under these circumstances, the probability of support is a lofty .5—hardly a rigorous experimental challenge.

In single-subject research, by comparison, treatment effects are clarified not by increasing statistical sensitivity but rather by improvements in experimental control. Individual differences are not averaged into obscurity as statistical error, but instead are regarded as revealing the limits of the control being exercised.

As a case in point, consider a situation encountered in the course of an experiment on "observing behavior" in adult humans (Perone & Baron, 1980). The main response was pulling a plunger mounted underneath a table. On the table was a console with colored stimulus lamps and several response keys. In the critical conditions, pressing the "observing" keys on the console would turn on colored lights correlated with the schedules of monetary reinforcement associated with the plunger response. During preliminary training, 1 subject adopted an unusual response topography: He tied one end of his bootlace to the plunger and the other end to the leg of his chair, put his feet on the table, and executed the response by rocking back and forth. When a monetary reinforcer was presented (an occasional event given the intermittent nature of the schedule), the subject repositioned himself and pressed a button on the console required to collect the reinforcer, then resumed the rocking motion. This topography was wholly compatible with the monetary schedule, which involved only the plunger response, but the investigators worried that it would interfere with the acquisition of the observing response, because the observing keys would usually be out of the subject's reach. To block the chair-rocking topography, the investigators replaced the chair with a wheeled stool. The subject reacted by sitting on the floor, tying his bootlace to the plunger and pulling the other end, and occasionally standing up to collect reinforcers. The new topography was no better than the old one. Finally, the investigators placed a limited hold on the collection button: Once a monetary reinforcer was earned, the subject had just 1 s to get up and collect it before it was canceled. This contingency was effective in moving the subject onto the stool in front the console, with the collection button and the observing keys within easy reach. When the critical phase of the experiment finally commenced, the subject acquired the observing response and his data fell in line with those of the other subjects.

The close interaction between investigator and subject fostered by the single-subject approach allowed a potential disaster to be identified and averted. The troublesome individual difference was not relegated to a statistical error term, but was eliminated by suitable adjustment in the experimental procedure. What would have happened in a group experiment? Perhaps the absence of a conditioned reinforcement effect in the problem subject would have been overlooked, if it did not appreciably affect the group mean. Or, if detected, the negative result might have been attributed to the regrettable but inevitable appearance in the sample of a recalcitrant subject whose person-

ality leads to sabotaging experimental goals. Of course, nothing about group designs prevents the kind of corrective action taken in this case. But the ready acceptance of individual differences and other forms of "error variance" seems more amenable to the theory and practice of group-statistical research than to single-subject research.

As noted elsewhere (Perone, 1991), a successful experimental science is one that exerts high degrees of control over its subject matter. The ability to control variables that affect behavior is prerequisite to the study of steady states. Thus, because single-subject designs require investigators to seek strict levels of control, their adoption encourages the development of an experimental science of behavior.

Judgment

Although some discussions of group-statistical methods may suggest otherwise, human judgment is an unavoidable component of the scientific enterprise. Investigators must exercise their best judgment repeatedly over the course of a research project. At the outset they must decide what line of investigation is likely to make a contribution to knowledge. Then they must devise appropriate experimental designs and procedures, often balancing competing interests based on convenience, economy, and the availability of apparatus and personnel. They must puzzle over the measures to employ, analyses to conduct, which results are worth reporting, and the implications of the results for contemporary theoretical debate. They must decide how methods, results, and arguments should be conveyed to the scientific community in the form of grant applications, publications, and professional presentations. Sometimes they must decide whether a negative outcome should spur a reappraisal of one's experimental strategy or abandonment of a cherished theoretical position. All these judgments and more are a matter of routine for active scientists regardless of their discipline, theoretical predilections, or epistemological convictions.

The adoption of group-statistical methods does not eliminate the need for an investigator's sound judgment, nor does the adoption of single-subject methods guarantee it. The two kinds of methods do, however, place different judgmental burdens on the investigator. And because of the relative rarity of single-subject methods, the burdens of that tradition are often misunderstood. Perhaps the greatest misunderstanding revolves around the so-called "visual analysis" of data.

When it comes to analyzing experimental results, the difference between group-statistical and single-subject methods is sometimes characterized along these lines: In group research, inferences about causal relations between independent and dependent variables are guided by precise, sophisticated statistical tests free of subjectivity and bias. In single-subject research, investigators stumble along with only a simple graph of the results to inspect unaided, leaving their causal inferences susceptible to all manner of idiosyncratic influences. Again, the absence of codified rules for conducting the visual analysis is seen as the culprit. Kazdin (1982) expressed the problem this way:

> Perhaps the major issue pertains to the lack of concrete decision rules for determining whether a particular demonstration shows or fails to show a reliable effect. The process of visual inspection would seem to permit, if not actively encourage, subjectivity and inconsistency in the evaluation of intervention effects. (p. 239)

Research is available to bolster this criticism. Investigators given session-by-session graphs of concocted behavioral data and asked to judge the presence of treatment effects may disagree with one another, be swayed by seemingly minor details of the graphic presentation, overlook small but reliable effects, or see effects when they are absent (DeProspero & Cohen, 1979; Knapp, 1983; Matyas & Greenwood, 1990; but see Parsonson & Baer, 1992, for a more appreciative account of visual analysis).

The rejoinder is that criticism of visual analysis is based on a profound misunderstanding. Indeed, the very term *visual analysis*—and the research into it—does not adequately represent the process as it occurs in actual research. Perhaps the problem can be traced to the comparison with statistical analysis. Statistical tests are conducted after an experiment is completed and the results are in. At that point, the investigator is left to sift though the data and seek evidence that an effect was brought about by the experimental manipulations. Critics of visual analysis seem to believe that it is merely an unsophisticated version of the same process: After the experiment the investigator draws a graph of the results and decides about the influence of the independent variable. But in practice no single-subject experiment is conducted in such a fashion. Visual analysis is an ongoing activity throughout the experiment; indeed, it is an integral part of the experimental analysis and as such it cannot be separated from the methods employed to collect the data in the graphs.

The point may be clarified by restating it with a more appropriate emphasis: *Experimental* analysis is an integral part of visual analysis. By this account, it is a mistake to suggest that investigators in the single-subject tradition prefer the visual inspection of graphs over statistical analysis. What *is* preferred is an experimental analysis so thorough, so powerful in its control over the subject matter of interest, that cause–effect relations are plain to see. The experiment may be regarded as any other scientific instrument, such as a microscope, whose resolution is painstakingly refined until the object of study comes into clear focus. The behavior analyst does not rely on unaided senses to see causal relations in behavior any more than the biologist relies on the naked eye to see subcellular objects. The adequacy of visual analysis depends on, and can be no greater than, the adequacy of the instrument aiding the investigator's vision, and in the study of behavior the instrument is the experiment. To be valid, a single-subject experiment must show that behavioral states can be replicated at will in different subjects and at different times within the same subject. Replication thus establishes the investigator's success in identifying and controlling relevant variables and confirms the adequacy of the stability criteria that guide the investigator's decisions about the attainment of steady states (see Baron & Perone, 1998, and Perone, 1991, for detailed discussion of the validity of single-subject experiments). In this connection, it is noteworthy that the previously cited research questioning the adequacy of visual analysis does not address the role of replication across subjects, nor does it express doubt about the conclusions of actual single-subject research.

Conclusions

The question prompting this essay is the role inferential statistics should play in behavior analysis. Ever since group-statistical methods gained favor in psychology, behavior analysis has drawn criticism for its devotion to single-subject methods. This essay has tried to show that the criticisms are based on an exalted and erroneous view of the power of statistical inference, one that regards statistical tests as a set of tried and true rules that reliably and inevitably guide investigators to objective answers for their experimental questions. In practice, however, statistical inference is not so simple. The rules, such as they are, have proven difficult to apply, even in the hands of statisticians, and the underlying logic of null-hypothesis testing has drawn fire since its popularization by Fisher nearly 75 years ago. Paradoxically, the criticism most often leveled against single-subject methods—that they do not ensure consistent outcomes across investigators—seems to apply equally to group-statistical methods.

Tests of statistical inference may

have their place in psychology, and perhaps even in behavior analysis. But there is no room for the unthinking methodological orthodoxy that often accompanies statistical inference. Perhaps the trouble started when Campbell and Stanley (1963) proclaimed that the only "true experiment" is one with random assignment of subjects to treatment groups. Campbell and Stanley directed their monograph to field researchers in education, and it seems unlikely that they intended to dismiss single-subject experiments (or, for that matter, virtually all natural science before 1925) as invalid. But by parroting Campbell and Stanley's monograph with insufficient thought or circumspection, several generations of textbooks on psychological research methods have surely had that unfortunate effect.

Whatever methods are adopted by behavior analysts, let us ask that they be adopted thoughtfully. The cookbook recipes sometimes associated with statistical inference are easy to criticize, but more thoughtful statistical applications may be welcome. In the same vein, it must be recognized that the demand for cookbooks is not altogether absent from the behavior-analytic community. Sidman (1960), as he wrote his *Tactics,* was perhaps the first to feel the demand. His response was to steadfastly refuse to offer any recipes. Instead, he asked his readers to think analytically about their research questions, to explore new procedures, and to learn from experience—in short, to develop good experimental judgment.

The present view, derived from the insights and advice offered by Sidman and Skinner, is that in a science of behavior good judgment is shaped by intensive interplay between investigator and subject in the course of experimental analysis. Group-statistical methods seem ill suited to the task, tending to insulate the investigator from the immediate results of experimental operations and reducing the motivation for seeking and exercising strong forms of control. By comparison, single-subject methods put investigator and subject into repeated contact, and force the investigator to identify and control variables relevant to the object of study. Thus, the methods are ideal for shaping—and maintaining—the kind of experimental practices that will ensure the continued success of behavior analysis.

REFERENCES

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 423–437.

Baron, A., & Perone, M. (1998). Experimental design and analysis in the laboratory study of human operant behavior. In K. A. Lattal & M. Perone (Eds.), *Handbook of research methods in human operant behavior* (pp. 45–91). New York: Plenum.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

DeProspero, A., & Cohen, S. (1979). Inconsistent visual analysis of intrasubject data. *Journal of Applied Behavior Analysis, 12,* 573–579.

Dixon, P. (1998). Why scientists value *p* values. *Psychonomic Bulletin & Review, 5,* 390–396.

Fisher, R. A. (1925). *Statistical methods for research workers.* Edinburgh, UK: Oliver & Boyd.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52,* 15–24.

Hagen, R. L. (1998). A further look at wrong reasons to abandon statistical testing. *American Psychologist, 53,* 801–803.

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings.* New York: Oxford University Press.

Knapp, T. (1983). Behavior analysts' visual appraisal of behavior change in graphic display. *Behavioral Assessment, 5,* 155–164.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70,* 151–159.

Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23,* 341–351.

McGrath, R. E. (1998). Significance testing: Is there something better? *American Psychologist, 53,* 796–797.

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34,* 103–115. (Reprinted in D. E. Morrison & R. E. Henkel, Eds.,

The significance test controversy, pp. 252–266. Chicago: Aldine, 1970)

Morrison, D. E., & Henkel, R. E. (Eds.). (1970). *The significance test controversy.* Chicago: Aldine.

Parsonson, B. S., & Baer, D. M. (1992). The visual analysis of data, and current research into the stimuli controlling it. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 15–40). Hillsdale, NJ: Erlbaum.

Perone, M. (1991). Experimental design in the analysis of free-operant behavior. In I. H. Iversen & K. A. Lattal (Eds.), *Techniques in the behavioral and neural sciences: Vol. 6. Experimental analysis of behavior, Part 1* (pp. 135–171). Amsterdam: Elsevier.

Perone, M., & Baron, A. (1980). Reinforcement of human observing behavior by a stimulus correlated with extinction or increased effort. *Journal of the Experimental Analysis of Behavior, 34,* 239–261.

Sidman, M. (1960). *Tactics of scientific research.* New York: Basic Books.

Skinner, B. F. (1938). *The behavior of organisms.* New York: Appleton-Century.

Skinner, B. F. (1972). A case history in scientific method. In B. F. Skinner (Ed.), *Cumulative record* (3rd ed., pp. 101–124). New York: Appleton-Century-Crofts. (Original work published 1956)

Thompson, B. (1998). In praise of brilliance: Where that praise really belongs. *American Psychologist, 53,* 799–800.

Tryon, W. W. (1998). The inscrutable null hypothesis. *American Psychologist, 53,* 796.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53,* 300–314.